Examples of Job-Time Filesystem Related Directives

This article provides a few examples of how the job-time related CLOUD directives are used.

Please review the following articles before you begin using the examples shown below:

- Creating a job-time filesystem
- Transferring files between S3 and a job-time filesystem
- Optional job-time filesystem related directives

Using Local Job-Time Filesystems

Data written to a local disk of a node are not accessible to processes running on other nodes. One scenario where the usage of such local disks is appropriate: running a CFD application where its MPI processes write out one file per rank or per node and the file is not needed by other ranks or nodes. The following examples walk you through this scenario.

Examples 1 and 2 show how to utilize the job-time local disks (/data, shown in the examples) for staging in the executable file (a.out) and input file (input.dat) from the Pleiades /nobackup system; and how to store rank-identified restart files generated by individual ranks in separate node-identified sub-folders of AWS S3 at the end of a job.

In Example 1, the ephemeral volume type is used when the local space needed for the restart file is less than 1.8 TB. In Example 2, the local volume type is used when the space needed is more than 1.8 TB.

In Example 3 (after you have followed the steps in Example 1 or Example 2), the restart files are loaded back from the S3 sub-folders to the local disk to resume simulation.

Note: If you do not store the restart files into sub-folders according to node numbers, when it is time to load the files back for the next job, each node will get all the restart files instead of just the ones needed for the ranks on the node.

Example 1

Specifying arch=c5d allows c5d.18x instances with 1.8 TB of NVMe SSD to be used as local disks.

```
#PBS -l select=10:mpiprocs=36:arch=c5d
...
#CLOUD -volume_type=ephemeral
#CLOUD -volume_put=outputfolder/{node}
#CLOUD -volume_mount=/data
..
#CLOUD -stagein_file=a.out
#CLOUD -stagein_file=input.dat
..

cd /data/

mpiexec -np 360 ./a.out < input.dat
```

Example 2

If you space needs more than 1.8 TB of space for the restart files, then specify the local volume type and a large volume size (for example, 2,048 GB).

```
#PBS -l select=10:mpiprocs=36:arch=c5d
...
#CLOUD -volume_type=local
#CLOUD -volume_put=outputfolder/{node}
#CLOUD -volume_mount=/data
#CLOUD -volume_size=2048
..
#CLOUD -stagein_file=a.out
#CLOUD -stagein_file=input.dat
..

cd /data/

mpiexec -np 360 ./a.out < input.dat
```

Example 3

Since the **-volume_put** directive was included when the previous job was done (as shown in examples 1 and 2), the files got saved off for this follow-on job to start from.

The **-volume_get** directive will bring back the files into the corresponding node in this new job. You do not need to include **{node}** in the **get** directive.

```
#PBS -l select=10:mpiprocs=36:arch=c5d
...
#CLOUD -volume_type=ephemeral
#CLOUD -volume_get=outputfolder
#CLOUD -volume_put=outputfolder
#CLOUD -volume_mount=/data
..
#CLOUD -stagein_file=a.out
#CLOUD -stagein_file=input.dat
..
cd /data/

mpiexec -np 360 ./a.out < input.dat
```

Using Shared Job-Time Filesystems

Example 4 shows how to create a filesystem (/nobackup, in this example) that only exists for the duration of the job and is shared by all nodes.

Example 4

```
#PBS -l select=10:mpiprocs=36
...
#CLOUD -volume_type=shared
#CLOUD -volume_size=2048
#CLOUD -volume_mount=/nobackup
..

cd /nobackup

mpiexec -np 360 ./a.out < input.dat
```

The directives **-volume_put** and **-volume_get** can be used if you want to save files in S3 between runs.

Example 5

If your job is such that the head node/rank does most of the writing, but there are some files that all ranks need to see, the **headnode_shared** volume type is useful.

```
#PBS -l select=10:mpiprocs=36
..
#CLOUD -volume_type=headnode_shared
#CLOUD -volume_size=2048
#CLOUD -volume_mount=/nobackup
...
```

Using Both Shared and Local Job-Time Filesystems

Example 6

Certain jobs may have some ranks that perform CFD computations and some that handle I/O or mesh regridding and need local disk space. You can use a shared filesystem for the computation and the node=X volume for those ranks that need local disk space. For example, if rank 10 and rank 50 need space but not the others, you can use:

```
#PBS -l select=10:mpiprocs=10
...
#CLOUD -volume_type=shared
#CLOUD -volume_size=100
#CLOUD -volume_mount=/nobackup
...
#CLOUD -volume_type=node=1
#CLOUD -volume_size=512
#CLOUD -volume_mount=/data
..
#CLOUD -volume_type=node=5
#CLOUD -volume_size=512
#CLOUD -volume_mount=/data
```

In this case, only node 1 and node 5 will have extra space in /data. The other nodes will use the shared filesystem, /nobackup.

Using a Non-Default Startup Script

Example 7

A job startup script under $HOME is useful to provide a customized environment for a login session or for different nodes in a PBS job. The customized environment may include the module load commands or set certain environment variables.

If the job does not use the persistent /home shared filesystem, where a startup script may be available, you can provide one that is accessible from the job's $PBS_O_WORKDIR, to be used in the place of the $HOME startup script. You can tell PBS to use that file instead of the system default ~/.bashrc file created on each compute node at boot time, as follows:

```
...
#CLOUD -volume_home_bashrc=my_bashrc_file
...
```

Article ID: 594
Last updated: 25 Jan, 2021
Revision: 38
Cloud Computing -> AWS Cloud -> Examples of Job-Time Filesystem Related Directives
https://www.nas.nasa.gov/hecc/support/kb/entry/594/